The principle of empirical risk minimization: mining for data stable patterns

Z.M. Shibzukhov[a], D.P. Dimitrichenko[a], M.A. Kazakov[a]

[a]*Institute of Applied Mathematics and Automation, branch of the Scientific Center of RAS, 360000, Shortanova street 89a, Nalchik, Russia*

Abstract

In this paper, we propose an extension for empirical risk minimization to solve the problem of regression. We applied averaging aggregation functions instead of the arithmetic mean to calculate the empirical risk. Such an intermediate risk assessment can be constructed using aggregate functions. These functions promote the solution to the problem for the penalty function minimization resulted from a deviation of its mean value. Such an approach to represent the aggregate average functions allows, on the one hand, to identify a much wider class of functions with mean-values.

In this paper we propose a new gradient scheme for solving the problem of minimizing the average risk. It is an analog circuit used in the SAG algorithm in the case when the risk is calculated with the arithmetic mean. Herein we present an illustrative example of the robust parameter estimation design in a linear regression based on the average function that approximates the median.

Keywords: aggregation function/operation; empirical risk; regression; penalty function; gradient descent

1. Introduction

Empirical Risk Minimization Method [1] is a recognized method for solving parametric regression. The empirical risk is usually calculated as the arithmetic mean of the parametric values of the loss function. Empirical assessment of average losses (arithmetical mean) is appropriate from a statistical point of view, if the losses are normally distributed. However, even applying the regular law the arithmetic mean is not a robust estimate of the average value. Meanwhile the median provides assessment of the observed mean value if emissions occur [2]. Therefore, empirical median estimator is also used in construction of parametric regression relations though it makes parameter setting slower [3]. When the emissions occur quantile estimators are also used if the distortion in losses distribution is less than 50%. This allows at parameters setting using median to save valuable part of the loss distribution. This median should be higher than the median that divide in ascending order a set of losses in two equal parts.

2. Empirical risk: the traditional approach

The problem of finding a parametric regression $y = f(\mathbf{x},\mathbf{w})$ between the input and scalar output y is one of the study problems of machine learning. There is a finite set of inputs $\tilde{\mathbf{X}} = \{\tilde{\mathbf{x}}_k : k = 1..\mathrm{N}\}$ and a set of known output values: $\tilde{\mathbf{Y}} = \{\tilde{y}_k : k = 1..\mathrm{N}\}$. A set of parameters $\mathbf{w}^*$ is to be such that $f^*(\mathbf{x}) = f(\mathbf{x},\mathbf{w}^*)$ to adequately represent the relationship between $\mathbf{x}$ and y in $\tilde{\mathbf{X}}$. An adequacy of f^* is frequently measured by empirical risk. Set parameters of $\mathbf{w}^*$ that specify adequate parametric relationship, should minimize the empirical risk. Empirical risk is usually calculated as the arithmetic mean of the parameters of the loss function:

$$\mathrm{Q}(\mathbf{w}) = \frac{1}{\mathrm{N}}\sum_{k=1}^{\mathrm{N}} \ell_k(\mathbf{w})$$

where $\ell_k(\mathbf{w}) = \ell_k(r_k(\mathbf{w}))$, where $\ell(r)$ is the loss function, $r_k(\mathbf{w}) = r\left(f(\tilde{\mathbf{x}}_k,\mathbf{w}),\tilde{y}_k\right)$ is a residual function between the value of the function *f* and the expected value at the *k*-th point. For example:

- the difference: $r(\mathbf{w}) = f(\mathbf{x},\mathbf{w}) - y$;
- the absolute difference: $r(\mathbf{w}) = \left|f(\mathbf{x},\mathbf{w}) - y\right|$;
- the asymmetric absolute difference: $\left|f(\mathbf{x},\mathbf{w}) - y\right|_\alpha$, where

$$|r|_\alpha = \begin{cases} \alpha r, & \text{if } r \geq 0 \\ (\alpha - 1)r, & \text{if } r < 0; \end{cases}$$

- the relative difference: $r(\mathbf{w}) = \frac{1}{|y|}|f(\mathbf{x},\mathbf{w}) - y|$ provided that values of y are separated from zero, or $(\mathbf{w}) = \frac{1}{1+|y|}|f(\mathbf{x},\mathbf{w}) - y|$.

Loss function is a non-negative function, which has a unique minimum such that $\ell(0) = \min \ell(r) = 0$. For example:

- absolute: $\ell(r) = |r|$;
- quadratic: $\ell(r) = r^2$;
- the Huber: $\ell(r) = \begin{cases} c(2|r| - c), \text{if } |r| \geq c \\ r^2, \text{ if } |r| < c; \end{cases}$
- the Tukey: $\ell'(r) = \begin{cases} r\left(1 - \left(\frac{r}{c}\right)\right)^2, \text{if } |r| \leq c \\ 0, \qquad \text{if } |r| > c; \end{cases}$
- the asymmetric absolute: $\ell(r) = \begin{cases} \alpha r, \text{ if } r \geq c \\ (\alpha - 1)r, \text{if } r < c; \end{cases}$
- the asymmetric quadratic: $\ell(r) = \begin{cases} \alpha r^2, \text{ if } r \geq c \\ (1-\alpha)r^2, \text{if } r < c; \end{cases}$

here $c > 0, 0 < \alpha < 1$.

From the statistical point of view, the evaluation of losses with the arithmetic mean is adequate, if the losses are normally distributed. However, if the losses are actually distributed according to the different law, the assessment of average losses should be carried out differently. But even in the case when losses are normally distributed the arithmetic mean is not stable in relation to emissions at the empirical distribution. In this case, a much more reasonable estimate is, for example, the median or quantile. The arithmetic mean, median, and quantiles are examples of averaging aggregation function. Therefore, in general, the average loss can be calculated using averaging aggregation functions [4,5].

3. From the empirical to the aggregate risk

Averaging aggregation functions have already been used in [8.9] for the construction of functional loss to operate the classification and regression algorithms. These algorithms retain the correctness property. We apply them now to estimate the average loss:

$$\mathrm{Q}_p(\mathbf{w}) = \mathrm{M}_p\left\{\ell_k(\mathbf{w}) : k = 1, \ldots, \mathrm{N}\right\}$$

where averaging aggregation function M_p is determined using the penalty function

$$\mathrm{M}_p\left\{\ell_1(\mathbf{w}), \ldots, \ell_{\mathrm{N}}(\mathbf{w})\right\} = \arg\min_u \sum_{k=1}^{N} p\left(\ell_k(\mathbf{w}), u\right)$$

The optimal set of parameter w^* supplies with minimum $\mathrm{Q}_p(\mathrm{w})$:

$$\mathrm{Q}_p(\mathbf{w}^*) = \min_u \mathrm{Q}_p(\mathbf{w})$$

If $p(z,u)$ contain partial derivatives up to the second order, then

$$\frac{\partial \mathrm{M}_p\{z_1,\ldots,z_n\}}{\partial z_k} = -\frac{p''_{uz}(z_k,\bar{z})}{\sum_{l=1}^{\mathrm{N}} p''_{uu}(z_k,\bar{z})}$$

where $\bar{z} = \mathrm{M}_p\{z_1,\ldots,z_{\mathrm{N}}\}$. Then

$$\operatorname{grad} \mathrm{M}_p\{\ell_1(\mathbf{w}),\ldots,\ell_N(\mathbf{w})\} = \frac{\sum_{k=1}^{N} - p''_{uz}(\ell_k(\mathbf{w}),\bar{z})\operatorname{grad}\ell_k(\mathbf{w})}{\sum_{k=1}^{N} p''_{uu}(\ell_k(\mathbf{w}),\bar{z})}$$

Search for the optimal set $\mathbf{w}$ can be carried out by the full gradient procedure. Updating of the parameters vector has the form of:

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - h_t \operatorname{grad} \mathrm{M}_p\{\ell_1(\mathbf{w}_{\mathrm{t}}),\ldots,\ell_N(\mathbf{w}_{\mathrm{t}})\}$$

Parameters vector is updated until values of $\mathbf{w}_t$ and $\mathrm{M}_p\{\ell_1(\mathbf{w}_{t+1}),\ldots,\ell_N(\mathbf{w}_{t+1})\}$ stabilize.

Note that if $p(z,u) = \mathrm{G}(z-u)$ is a special case of the general equation

$$p(z,u) = \mathrm{G}(h(\mathrm{z}) - h(u))$$

where $\mathrm{G}: \mathrm{R} \to \mathrm{R}$ is strictly continuous convex function, $h(u)$ is strictly monotonic function [7, 6].

$$\operatorname{grad} \mathrm{M}_p\{\ell_1(\mathbf{w}),\ldots,\ell_{\mathrm{N}}(\mathbf{w})\} = \sum_{k=1}^{N} \alpha_k(\mathbf{w}) \operatorname{grad} \ell_k(\mathbf{w}) \quad (1)$$

where

$$\alpha_k(\mathbf{w}) = \frac{\mathrm{G}''(\ell_k(\mathbf{w}) - \bar{z})}{\mathrm{G}''(\ell_1(\mathbf{w}) - \bar{z}) + \ldots + \mathrm{G}''(\ell_{\mathrm{N}}(\mathbf{w}) - \bar{z})},$$

and $\alpha_1(\mathbf{w}) + \ldots + \alpha_{\mathrm{N}}(\mathbf{w}) = 1$.

It is easy to see that in this case, the gradient descent algorithm seems similar to search algorithms for minimizing average loss weighted with the numeric weights. However, here the weights are functions of $\ell_1(\mathbf{w}) - \bar{z}, \ldots, \ell_{\mathrm{N}}(\mathbf{w}) - \bar{z}$ deviations between the aggregate losses and current losses. If $\mathrm{G}(z-u) = (z-u)^2/2$, then $\alpha_k(\mathbf{w}) = 1/\mathrm{N}$, which corresponds to the arithmetic mean of the loss or value of the empirical risk.

The pseudo-code of setting algorithm for $\mathbf{w}$ based on the total gradient method. PBFG algorithm.

From the computational point of view the mentioned algorithm is not optimal, since on each iteration step to calculate the aggregate average it is necessary to find the minimum value of the function. Therefore let us consider another iterative algorithm that searches for values of $\mathbf{w}^*$ and $\mathrm{M}_p\{\ell_1(\mathbf{w}^*),\ldots,\ell_{\mathrm{N}}(\mathbf{w}^*)\}$ simultaneously.

Algorithm PBFG: Complete gradient descent algorithm based on aggregate function.

$t \leftarrow 0$

Initialize $\mathbf{w}_0$

$u_0 \leftarrow \mathrm{M}_p\{\ell_1(\mathbf{w}),\ldots,\ell_{\mathrm{N}}(\mathbf{w}_0)\}$

repeat

$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - h_t \operatorname{grad} \mathrm{M}_p\{\ell_1(\mathbf{w}_t),\ldots,\ell_{\mathrm{N}}(\mathbf{w}_t)\}$

$u_{t+1} \leftarrow \mathrm{M}_p\{\ell_1(\mathbf{w}_{t+1}),\ldots,\ell_1(\mathbf{w}_{t+1})\}$

$t \leftarrow t+1$

until $\{u\}$ and $\{\mathbf{w}_t\}$ stabilizes.

4. Stochastic average gradient algorithm based on aggregate functions

Since gradient (1) is a weighted sum of gradients of corresponding losses it is possible to apply a technique with SAG (Stochastic Average Gradient) [10,11]. We apply this technique to design PBSAG algorithm – Penalty Based Stochastic Average Gradient. A scheme for parameters adaptation $\mathbf{w}$ and u has the form of:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - h_t \overline{\mathbf{g}}_t .$$

$$u_{t+1} = u_t - \tau_t \overline{q}_t$$

where

$$\overline{\mathbf{g}}_t = \frac{\sum_{k=1}^{\mathrm{N}} \mathbf{g}_{k,t}}{\sum_{k=1}^{\mathrm{N}} g_{k,t}}$$

A value of $\overline{q}_t$ can be updated in accordance with one of the following rules:

$$\overline{q}_t = \frac{1}{\mathrm{N}} \sum_{k=1}^{\mathrm{N}} q_{k,t} \quad \text{or} \quad \overline{q}_t = \frac{\sum_{k=1}^{\mathrm{N}} q_{k,t}}{\sum_{k=1}^{\mathrm{N}} g_{k,t}}$$

whichever used gradient descent method or Newton's method to find the minimum value of the averaging aggregation function M_p. Vectors of the set $\left\{\mathbf{g}_{k,t} : k = \overline{1,\mathrm{N}}\right\}$ are updated by the following rule:

$$\mathbf{g}_{k,t+1} = \begin{cases} -p''_{uz}(\ell_k(\mathbf{w}_t), u_t)\operatorname{grad}\ell_k(\mathbf{w}_t), & \text{if } k = k(t) \\ \mathbf{g}_{k,t}, & \text{else.} \end{cases}$$

Values of $\left\{g_{k,t} : k = \overline{1,\mathrm{N}}\right\}$ and $\left\{q_{k,t} : k = \overline{1,\mathrm{N}}\right\}$ are updated in correspondence with the following rules:

$$g_{k,t+1} = \begin{cases} p''_{uu}\left(\ell_k(\mathbf{w}_t), u_t\right), & \text{if } k = k(t) \\ g_{k,t}, & \text{else,} \end{cases} .$$

$$q_{k,t+1} = \begin{cases} p'_{u}\left(\ell_k(\mathbf{w}_t), u_t\right), & \text{if } k = k(t) \\ q_{k,t}, & \text{else.} \end{cases}$$

Algorithm PBSAG: Stochastic average gradient algorithm based on the average function.

$t \leftarrow 0$
Initialize $\mathbf{w}_0$
for $k \in \{1,\ldots,\mathrm{N}\}$ **do**
$\mathbf{G}_k \leftarrow p''_{uz}\left(\ell_k(\mathbf{w}_0), u_0\right)\operatorname{grad}\ell_k(\mathbf{w}_0)$
$\mathrm{H}_k \leftarrow p''_{uu}\left(\ell_k(\mathbf{w}_0), u_0\right)$
$\mathrm{Q}_k \leftarrow p'_{u}\left(\ell_k(\mathbf{w}_0), u_0\right)$
end for
$\mathbf{G} \leftarrow \mathbf{G}_1 + \ldots + \mathbf{G}_{\mathbf{N}}$
$\mathrm{H} \leftarrow \mathrm{H}_1 + \ldots + \mathrm{H}_{\mathrm{N}}$
$\mathrm{Q} \leftarrow \mathrm{Q}_1 + \ldots + \mathrm{Q}_{\mathrm{N}}$
repeat
$k = k(t)$

$\mathbf{G} \leftarrow \mathbf{G} - \mathbf{G}_{\boldsymbol{k}} + p''_{uz}\left(\ell_k(\mathbf{w}_0), u_t\right)\operatorname{grad}\ell_k(\mathbf{w}_t)$

$\mathbf{G}_{\boldsymbol{k}} \leftarrow p''_{uz}\left(\ell_k(\mathbf{w}_0), u_t\right)\operatorname{grad}\ell_k(\mathbf{w}_t)$

$\mathrm{H} \leftarrow \mathrm{H} - \mathrm{H}_k + p''_{uu}\left(\ell_k(\mathbf{w}_0), u_t\right)$

$\mathrm{H}_{\boldsymbol{k}} \leftarrow p''_{uu}\left(\ell_k(\mathbf{w}_0), u_t\right)$

$\mathrm{Q} \leftarrow \mathrm{Q} - \mathrm{Q}_k + p'_u\left(\ell_k(\mathbf{w}_0), u_t\right)$

$\mathrm{Q}_{\boldsymbol{k}} \leftarrow p'_u\left(\ell_k(\mathbf{w}_0), u_t\right)$

$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - h_t\bar{\mathbf{g}}$

if Newton's scheme is used **then**

$\bar{q} \leftarrow \mathrm{Q}/\mathrm{G}_2$

else

$\bar{q} \leftarrow \mathrm{Q}/\mathrm{N}$

end if

$u_{t+1} \leftarrow u_t - \tau_t\bar{q}$

$t \leftarrow t+1$

until $\{u_t\}$ and $\{\mathbf{w}_t\}$ stabilize.

Algorithm PBSAG on each step should store one gradient vector and two values for each sample of the training data set, i.e. real numbers $\mathrm{N}(m+2)$, where m is grade of vector of $\mathbf{w}$. Therefore, it should be applied if there exists a memory for such data volume storage.

It is easily seen that if $p(z,u) = (z-u)^2/2$, then PBSAG algorithm scheme reduces to the SAG scheme:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - h_t\mathbf{g}_t$$

where

$$\mathbf{g}_t = \frac{1}{\mathrm{N}}\sum_{k=1}^{\mathrm{N}}\mathbf{g}_{k,t},\ \mathbf{g}_{k,t+1} = \begin{cases}\operatorname{grad}\ell_k(\mathbf{w}_k), & \text{if } k = k(t)\\ \mathbf{g}_{k,t} & \text{else.}\end{cases}$$

Thus PBSAG scheme is a regular extension of the SAG algorithm scheme [10,11], when to calculate the average losses averaging aggregation function based on the penalty is applied instead of the arithmetic mean.

5. PBSAG application examples

Let us consider the PBSAG application with «approximated» LMS to find linear regression under conditions of emissions. In the standard algorithm of LMS the least square error is searched:

$$\mathrm{E}(\mathbf{w}) = \operatorname{med}\left\{\left(f(\mathbf{x}_k,\mathbf{w}) - y_{\boldsymbol{k}}\right)^2 : \mathrm{k} = 1,\ldots,\mathrm{N}\right\}.$$

The PBSAG is not applied when M is the median. However, it can be replaced by its asymptotically equivalent substitute.

Definition. The expression $p_\alpha(z-u)$ defines the asymptotically equivalent median substitute provided that for some α^*

$$\lim_{\alpha\to\alpha^*} p_\alpha(z-u) = |z-u|;$$
$$\lim_{\alpha\to\alpha^*} p'_\alpha(z-u) = \operatorname{sign}(z-u)$$

Consider the example:

$$p_\alpha(z-u) = \sqrt{\alpha^2 + |z-u|^2} - \alpha,\ \text{when } \alpha^* = 0$$

$$p'_\alpha(z-u) = \frac{z-u}{\left(\alpha^2+|z-u|^2\right)^{1/2}}, \qquad p''_\alpha(u-z) = \frac{\alpha^2}{\left(\alpha^2+|z-u|^2\right)^{3/2}}.$$

We call an α -median a corresponding averaging aggregation function:

$$\mathrm{med}_{\alpha}\{z_1,\ldots,z_{\mathrm{N}}\} = \arg\min_{u}\sum_{k=1}^{\mathrm{N}}\left(\sqrt{\alpha^2+|z_k-u|^2}-\alpha\right).$$

Another substitute can be constructed using the following dissimilarity function

$$p_{\alpha}(z-u) = |z-u| - \alpha\ln\left(\alpha+|z-u|\right)+\alpha\ln\alpha \tag{2}$$

Figure 1 shows the application of the PBSAG algorithm. It shows the PBSAG method and algorithm capacity based on a true averaging aggregation functionality that approximate median (2) to restore a linear regression relations when the emissions is contained in the initial data.

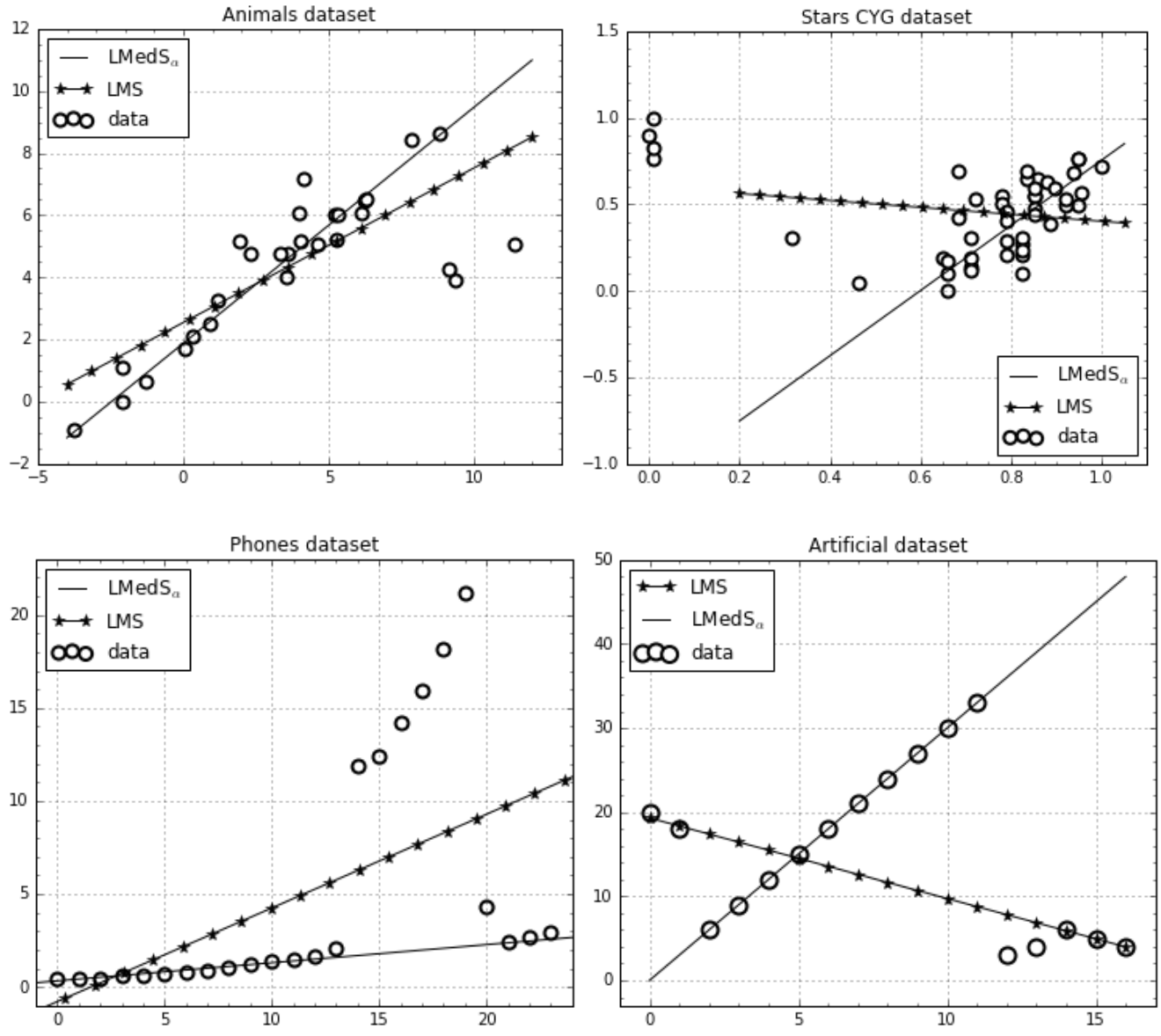

Fig. 1. Examples of linear regression recovery using the LMS method and LMedSα (α = 0.001).

6. Conclusion

The functions mentioned in this paper are used to estimate average loss. The averaging aggregation function is determined using the penalty function. As a result, the gradient descent becomes similar to the search procedure for the minimum of the weighted average of losses resulted from the numeric scales use. This makes construction of the PBSAG algorithm (Penalty Based Stochastic Average Gradient) based on the true averaging aggregation function possible. The proposed algorithm that implements the empirical risk minimization method allows to cope with the task of rebuilding of the linear regression relations when the emission is contained in the initial data. The algorithm property is illustrated with examples.

Acknowledgements

This work was supported by RFBR grant 15-01-03381 and (DNIT) RAS Division of Nanotechnologies and Information Technologies grant

References

[1] Vapnik, V. The Nature of Statistical Learning Theory / V. Vapnik – Information Science and Statistics. – Springer-Verlag, 2000.
[2] Rousseeuw, P.J. Least Median of Squares Regression / P.J. Rousseeuw // Journal of the American Statistical Association. – 1984. – No.79. – P.871–880.
[3] Rousseeuw, P.J. Robust Regression and Outlier Detection. / P.J. Rousseeuw, Leroy – New York: John Wiley and Sons, 1987.
[4] Mesiar, R. Aggregation functions: A revision. / R. Mesiar, M. Komornikova, A. Kolesarova, T. Calvo // H. Bustince, F. Herrera, J. Montero, editors, Fuzzy Sets and Their Extensions: Representation, Aggregation and Models. – Springer. Berlin. Heidelberg – 2008.
[5] Grabich, M. Aggregation Functions / M. Grabich, J.-L. Marichal, E. Pap // Series: Encyclopedia of Mathematics and its Applications, No.127 – Cambridge University Press. – 2009.
[6] Beliakov, G. A Practical Guide to Averaging Functions / G. Beliakov, H. Sola, T. Calvo – Springer – 2016. – 329 p.
[7] Calvo, T., Beliakov, G. Aggregation functions based on penalties / T. Calvo, G. Beliakov // Fuzzy Sets and Systems – 2010. Vol.161, No.10, PP.1420-1436.
[8] Shibzukhov, Z.M. Correct Aggregate Operations with Algorithms / Z.M. Shibzukhov // Pattern Recognition and Image Analysis – 2014. Vol. 24. No. 3. PP. 377–382.
[9] Shibzukhov, Z.M. Aggregation correct operations on algorithms / Z.M. Shibzukhov // Doklady Mathematics – 2015. Vol. 91. No. 3. PP. 391-393.
[10] Le Roux, N. A Stochastic Gradient Method with an Exponential Convergence Rate for Finite Training Sets / N. Le Roux, M. Schmidt, F. Bach // http://arxiv.org/abs/1202.6258
[11] Schmidt, M., Le Roux, N., Bach, F. Minimizing Finite Sums with the Stochastic Average Gradient / M. Schmidt, N. Le Roux, F. Bach // arXiv.org, 2013. http://arxiv.org/abs/1309.2388